DOCUMENT RESUME

ED 298 143 TM 012 229

AUTHOR Marsh, Herbert W.; And Others
TITLE Masculinity and Femininity: A Bipolar Construct and Independent Constructs.
PUB DATE 11 Jun 88
NOTE 39p.
AVAILABLE FROM Herbert W. Marsh, Faculty of Education, University of Sydney, New South Wales 2006, Australia.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Androgyny; College Students; Construct Validity; *Factor Analysis; Multitrait Multimethod Techniques; Personality Measures; Psychometrics; *Rating Scales; *Self Esteem; *Sex Differences; *Social Desirability
IDENTIFIERS Bipolar Trait Ratings Scales; Confirmatory Factor Analysis; *Masculinity Femininity Variable

ABSTRACT

Data from the authors' previous research (1979, 1980, 1987), consisting of responses to five masculinity-femininity (M-F), two esteem, and two social desirability instruments, were reanalyzed. The subjects were 104 male and 133 female college students who completed the: Bem Sex Role Inventory, Personal Attributes Questionnaire, ANDRO instrument and Social Desirability Scale from the Personality Research Form, Femininity Scale of the California Psychological Inventory, Masculinity versus Femininity Scale of the Comrey Personality Scales, Feelings of Inadequacy Scale, and Self-Acceptance Scale. Correlations between M and F for the five M-F instruments varies from 0.23 to about -1.0. Support for distinguishable (non-bipolar) M and F factors was found for four of the instruments. Applying confirmatory factor analysis (CFA) and hierarchical CFA, the dimensionality of M-F and the influence of method/halo effects in responses to specific instruments were studied. The best fitting model identified three higher-order factors. In support of traditional personality theories, one factor was a bipolar M-F construct, but in support of androgyny theory, the other two factors were distinguishable M and F factors. The factor structures were reasonably invariant for men and women; methodological implications of this finding are substantial. In subsequent analyses, the higher-order M-F factors were related to esteem, social desirability, and gender in order to further test interpretations of the M-F factors. Six tables and four figures conclude the document. (TJH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED298143

Masculinity and Femininity: A Bipolar Construct and Independent Constructs

Herbert W. Marsh
University of Sydney, Australia

John K. Antill & John D. Cunningham
Macquarie University, Australia

Revised 11 June, 1988.

Running Head: Masculinity and Femininity

Requests for reprints should be sent to Herbert W. Marsh, Faculty of Education, The University of Sydney, NSW 2006, Australia. The authors would like to thank Jack McArdle and Jennifer Barnes for helpful comments on earlier drafts of the paper.

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

✓ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

HERBERT W. MARSH

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM012 229

Masculinity and Femininity:

A Bipolar Construct and Independent Constructs

ABSTRACT

The present investigation is a reanalysis of data from Antill and Cunningham (1979; 1980; Marsh, Antill & Cunningham, 1987) consisting of responses to five Masculinity-Femininity (MF) instruments, two esteem instruments, and two social desirability scales. Correlations between M and F for the 5 instruments varied from .23 to approximately -1.0; support for distinguishable (non-bipolar) M and F factors was found for 4 of the instruments. Applying confirmatory factor analysis (CFA) and hierarchical CFA (HCFA), the present study examined the dimensionality of MF and the influence of method/halo effects in responses to specific instruments. The best fitting model identified three higher-order factors; in support of traditional personality theories one factor was a bipolar MF construct, but in support of androgyny theory the other two factors were distinguishable M and F factors. The factor structures were reasonably invariant for men and women, and methodological implications of this important finding were examined. In subsequent analyses, the higher-order MF factors were related to esteem, social desirability, and gender in order to further test interpretations of the MF factors.

BEST COPY AVAILABLE

Masculinity and Femininity:
A Bipolar Construct and Independent Constructs

The present investigation is a reanalysis of data from Antill and Cunningham (1979; 1980; Marsh, Antill & Cunningham, 1987) consisting of responses to five Masculinity-Femininity (MF) instruments, two Self esteem instruments, and two Social Desirability instruments. The purposes of the present investigation are to: a) examine the dimensionality of MF; b) examine the relation of derived MF factors to other constructs (esteem, social desirability, and gender); c) demonstrate the implications of testing the dimensionality for men and women separately, for the total-group covariance matrix, or for the pooled within-group covariance matrix that removes the effect of gender; and d) demonstrate recent advances in the application of confirmatory factor analysis (CFA) and hierarchical CFA (HCFA) to such problems.

The MF Construct and Its Relation to Esteem and Social Desirability

The Dimensionality of MF

Virtually all researchers prior to Constantinople's 1973 review and many current personality inventories assume that M and F are the end-points of a single, bipolar dimension. This implies that the correlation between M and F is close to -1.0. More recently, androgyny researchers have argued that it is logically possible to be both M and F, and the existence of both in the same person has been labeled androgyny. The two key hypotheses of androgyny theory are that: (a) M and F are distinguishable dimensions and (b) individuals high on both M and F are mentally healthier and socially more effective. A considerable and growing body of research has been directed at contrasting these two apparently opposing views of M and F (see Marsh & Myers, 1986, for a review).

In support of androgyny theory, androgyny researchers have typically found that MF correlations (i.e., correlations between M and F) differ significantly from -1.0. However, Marsh and Myers (1986) found that MF correlations for different instruments varied from moderately positive to close to -1.0. They showed how such differences were logically consistent with the design of the instruments. For example, the use of only socially desirable attributes to represent M and F may produce a response bias that results in a near-zero or positive MF correlation that is consistent with androgyny theory. Alternatively, the use of logically opposed items to represent M and F is likely to result in a substantially negative MF correlation that is consistent with a bipolar MF. Also, in an exploratory

factor analysis of the original BSRI items, Pedhauzur and Tetenbaum (1979) found that responses to the adjectives "masculine" and "feminine" were substantially negatively correlated and formed a two-item bipolar factor. In their factor analysis they reported four orthogonal factors defined by traditionally feminine items, traditionally masculine items, the bipolar MF factor defined by the "masculine" and "feminine" adjectives and an additional factor which they called self-sufficiency. On the basis of this four-factor solution, they concluded that: "The fact that the traits Masculine and Feminine describe a separate bipolar factor also casts doubt on the validity of the classification of the remaining items as masculine and feminine" (p. 1012). Whereas most research such as that summarized here has sought to establish the structure of MF in separate analyses of individual instruments, the purpose of the present investigation is to establish the structure of MF across responses to five different instruments.

MF as Measured on the Five MF Instruments Used Here.

The five MF instruments considered in the present investigation represent very different approaches to the measurement and conceptualization of the MF construct. The MF scale from the California Psychological Inventory (CPI; Megargee, 1972), like many traditional personality instruments, contains items that maximally differentiate men and women so that M and F scores are highly correlated with gender. Because biological gender is bipolar, this type of scale is likely to also be bipolar. In an alternative approach, the Comrey Personality Scales (CPS; Comrey, 1970; also see Marsh, 1985) is based on distinct item clusters designed to represent components of MF on a logical/theoretical basis that were substantiated by factor analysis. Consistent with the CPS assumption of bipolarity, logically opposed items were used to reflect the M and F endpoints within each of the item clusters. Hence both these traditional personality instruments conceptualize MF to be a bipolar construct.

The Bem Sex Role Inventory (BSRI; Bem, 1974) is based on socially desirable items empirically rated to be more desirable for one sex or the other. In contrast, the Personal Attributes Questionnaire (PAQ; Spence, 1984) is based on socially desirable items rated to be more typical of one sex or the other. Spence and Bem also offer theoretical distinctions between the two instruments such as the generality of the M and F constructs inferred by the two instruments. Spence (1984) emphasized that PAQ measures two trait clusters that can be labeled dominance/self-assertiveness (PAQ M) and nurturance/interpersonal orientation (PAQ F). Nevertheless, both PAQ and

BSRI are based on socially desirable attributes, both result in distinguishable (non-bipolar) constructs, and PAQ scores are highly correlated with BSRI scores (Cook, 1985). Frequent criticisms of these instruments include: a) their reliance on socially desirable attributes; b) their atheoretical approach to instrument construction; and c) the limited scope of the M and F traits based on them (for further discussion see Cook, 1985; Kelly & Worell, 1977; Locksley & Colten, 1979; Marsh & Myers, 1986; Pedhazur & Tetenbaum, 1979; Spence, 1984). Nevertheless, these two scales continue to be the most frequently used in androgyny research. [The original version of PAQ also included a bipolar MF scale and subsequent versions of PAQ included M and F scales derived from negatively valued items, but these additional PAQ scales were not considered here.]

The ANDRO scale (Berzins, Welling & Wetter, 1978) was developed by selecting existing items from the Personality Research Form (PRF; Jackson, 1967) according to their sex-typed desirability and their consistency with the content themes in the BSRI. Ratings by university undergraduates were used to corroborate the sex-typed desirability of the items. The rationale for ANDRO was to develop an instrument consistent with the BSRI based on PRF responses so that the androgyny construct could be examined in the wide range of studies that have used, and will use, the PRF. Hence, the conceptualization of the ANDRO MF scales, though based on items from a traditional personality instrument, is similar to BSRI and PAQ.

The five MF instruments considered in the present investigation differ substantially in their conceptualization and design. Hence, an important question is the extent to which they measure common M and F traits. Multitrait-multimethod (MTMM) analysis (Campbell & Fiske, 1959; Marsh, in press; Marsh & Hocevar, 1983) is ideally suited to examine this question. Within this MTMM perspective there are two traits (M and F) and five methods (the five MF instruments). The substantive questions to be examined are: a) to what extent can the M and F scores from each instrument be combined to form a global M, a global F, or a bipolar global MF? b) what are the relations among these global measures? and c) what are the influences of method effects that are idiosyncratic to particular instruments?

Relations Between MF Responses and Other Constructs.

Esteem. Androgyny theory posits that both M and F, or perhaps the M x F cross-product, should contribute positively and uniquely to esteem. Extensive reviews have examined the MF/esteem relation and theoretical models of this relation (e.g., Hall & Taylor, 1985; Lubinski, Tellegen &

Butcher, 1983; Marsh, 1987; Marsh, Antill & Cunningham, 1987; Spence, 1984; Whitley, 1983). However, empirical findings indicate that whereas the contribution of M is substantial and positive, the unique contribution of F is nil or negative (but see Marsh, 1987) as is M x F. Furthermore, the effects of none of these variables seem to depend on gender. Marsh, Antill and Cunningham (1987) tested various models of the MF/esteem relation with the data considered here. They found that the unique contribution of M to esteem was consistently more positive than that of F which was either nil or negative, did not vary with gender as posited by sex-typed models, and did not interact with F as posited by interactive androgyny models.

Social desirability. Social desirability is an inferred response bias or method effect whereby individuals respond to the desirability an item instead of or in addition to the specific item content. Methodological issues related to social desirability are important in androgyny research (Marsh, 1987; Marsh, Antill & Cunningham, 1987). The MF correlation is probably influenced by the social desirability of the M and F items. If both M and F items are consistently positive, or consistently negative in terms of social desirability, then the MF correlation is likely to be more positive than if the M and F items are neutral. Furthermore, if M and F items are consistently high in terms of social desirability, then the MF/esteem relation may be explicable in terms of the social desirability of the MF items instead of their specific M or F content. Finally, if the social desirability of M items differs substantially from that of F items, then the differential influence of M and F to the prediction of desirable outcomes may be due to differences in social desirability instead of differences in the M and F content of the items.

The influence of social desirability is often viewed as an undesirable response bias or source of invalidity, but this view may be too simplistic. For example, high scores on both esteem and social desirability measures are typically inferred from positive responses to socially desirable attributes and negative responses to socially undesirable attributes so that esteem and social desirability responses should be substantially correlated. Marsh, Antill and Cunningham (1987) found esteem to be more positively correlated with M than F, social desirability was more positively correlated with F than M. They speculated that esteem items may be stereotypically more masculine whereas social desirability may be stereotypically more feminine. Consistent with this explanation, males had higher esteem scores than females, but females had higher social desirability scores than males.

Gender. Not surprisingly, males tend to have higher M scores than females whereas females tend to have higher F scores than males. The size of these sex differences, however, varies substantially depending on the MF instrument. Marsh and Myers (1986) found that responses to instruments designed to infer a bipolar MF construct were more substantially related to gender than were responses to instruments designed to measure independent M and F constructs. The relation of gender to M and F scales also complicates the examination of the factor structure of MF responses. Important questions to be addressed are whether factor structures are invariant for males and females and whether the influence of biological gender is a valid source of influence in the formation of M and F.

Method

Sample and Materials

The sample, materials, and the collection of data are described in more detail by Antill and Cunningham (1979; 1980). Briefly the subjects were 104 male and 133 female college students who completed: a) the Bem Sex Role Inventory (BSRI; Bem, 1974) consisting of 20 M, 20 F, and 20 neutral (Social Desirability) adjectives; b) the Personal Attributes Questionnaire (PAQ; Spence, 1984) consisting of 23 M and 18 F adjectives; c) the ANDRO instrument (Berzins, Welling & Wetter, 1978) consisting of 29 M and 27 F items from the Personality Research Form (PRF; Jackson, 1967) and the Social Desirability scale from the PRF; d) the Femininity scale of the California Psychological Inventory (CPI; Gough, 1957) consisting of 21 M items and 17 F items; e) the Masculinity versus Femininity scale of the Comrey Personality Scales (CPS; Comrey, 1970; also see Marsh, 1985) consisting of 10 M and 10 F items; f) the Feelings of Inadequacy Scale (Janis & Field, 1959, as revised by Eagly, 1967) consisting of 20 esteem items; and g) the Self-Acceptance Scale (Berger, 1952) consisting of 36 esteem items. The first three MF measures (BSRI, PAQ, ANDRO) were explicitly designed as androgyny measures and provide separate M and F scores. The CPI and CPS were designed to infer a bipolar MF, but separate M and F scores can be constructed by scoring M and F items separately.

Statistical Analyses.

Preliminary analyses. Psychometric properties of the self-report scales are summarized in Table 1. For all five MF instruments, M, F, and bipolar MF (M items scored positively and F items scored negatively) have at least modest coefficient alpha estimates of reliability and correlate substantially with gender in the expected direction. The bipolar MF from

the CPI that was originally devised to differentiate between males and females, and it correlates with gender at a level close to the reliability of the scale. MF correlations for the other instruments are smaller but still substantial. The correlations between M and F scales vary from modestly positive to close to -1.0. The negative correlation for the CPS approaches the reliability of its M and F scales in a manner that is consistent with its bipolar conceptualization of MF. Actually, after correction for unreliability the MF correlation for the CPS is slightly more negative than -1.0. Consistent with their design, the M and F scores for the BSRI and PAQ are positively correlated with social desirability. In contrast, social desirability is less positively correlated with the M and F scales from the CPS and the ANDRO, and negatively correlated with the CPI scales.

Insert Table 1 About Here

For all five MF instruments, M scores are more positively correlated with esteem than are F scores so that all five bipolar MF scores are positively correlated with esteem (Table 1). Though not reported here, other analyses of this data (Marsh, Antill and Cunningham, 1987) indicated that for all five MF instruments the contribution of F after controlling for M was nil or negative, the M x F crossproduct did not contribute to esteem beyond the contribution of M and F, and the effects of M, F and M x F did not interact with gender. Controlling for social desirability did not alter the general pattern of results, and all interaction effects -- M x F and those involving gender -- were still nonsignificant.

Confirmatory Factor Analyses (CFAs). For present purposes, items from each of the 14 (5 M, 5 F, 2 social desirability, and 2 esteem) scales were randomly divided into thirds (in subsequent discussion these are called item parcels or simply parcels). A covariance matrix derived from these 42 (14 x 3) parcels for all subjects was the basis of the CFAs. The large number of items in these 14 scales -- 251 -- precluded the analysis of responses to individual items. Furthermore, there are important advantages to analyzing responses to subscale scores instead of items: (a) parcel scores typically have greater reliability and generality, (b) response biases and other characteristics that are idiosyncratic to individual items are likely to have less influence, (c) the ratios of measured variables to inferred factors and to estimated parameters are increased, and (d) distributions of the measured variables are less likely to cause problems for factor analyses -- particularly when item responses are dichotomous.

Box's M (SPSS, 1986) was used to test the equality of the

variance/covariance matrices for males and females. In results discussed latter in more detail, the two covariance matrices did not differ significantly ($p > .05$) for the 30x30 matrix based on the 30 MF parcels derived from the 5 MF instruments, or the 42x42 matrix that included the 6 esteem and 6 social desirability parcels. Based in part on these findings, the focus of subsequent results was on CFAs conducted on the total group covariance matrix.

In preliminary, unreported CFAs, various one- and two-factor models were fit to responses from each instrument separately. These results indicated that two-factor (M and F) solutions fit responses to all but the CPS instrument substantially better than did one-factor solutions. Correlations between the M and F factors in the two factor solutions were similar to the MF correlations between scale scores that have been corrected for unreliability (Table 1). In the first set of analyses considered here, models were fit to responses from all five MF instruments. In subsequent analyses, relations between MF responses and other variables were considered. The CFAs were conducted with LISREL V (Joreskog & Sorbom, 1981). Introductions to the use of CFA and LISREL are available elsewhere (e.g., Bagozzi, 1981; Joreskog, 1981; Joreskog & Sorbom, 1981; Long, 1983; Marsh, 1985; Marsh & Hocevar, 1983; 1984; 1985; 1988; Pedhauzur, 1982) and so are not presented in detail. The details of these models are presented in the Results section (also see Appendix I).

In CFA there are not well-established guidelines for testing goodness of fit. The general approach, and the one used here, is to: a) examine parameters in relation to substantive issues; b) evaluate the overall X^2 goodness of fit in terms of statistical significance and in comparison to alternative models; and c) evaluate subjective goodness-of-fit indicators such as the X^2/df ratio and the Tucker Lewis Index (TLI; Marsh, Balla & McDonald, 1988) and to compare values from alternative models.

A related problem is the occurrence of Heywood cases, parameter estimates that are outside of the range of allowable values, such as residual variance estimates that are negative. Heywood cases are more likely when the sample size is small relative to the number of parameters that are estimated, when there are few indicators for each factor, and when the factor structure is complex (e.g., variables are associated with more than one factor). Heywood cases are likely to represent sampling error when the confidence interval about the improper parameter estimate contains proper values and the size of its standard error is reasonable (Gerbing & Anderson,

1987; Van Driel, 1978). For example, for simulated data tested with the true population model, Anderson and Gerbing (1984) found that 25% of the solutions contained Heywood cases. However, the occurrence of such Heywood cases had little effect on parameter estimates for other factors or on goodness-of-fit (Gerbing & Anderson, 1987). Hence, improper parameter estimates are unlikely to substantially affect substantive conclusions as long as confidence intervals about improper estimates contain proper values, and standard errors are reasonable. In alternative approaches to this problem (e.g., Dillon, Kumar & Mulani, 1987) it is possible to artificially restrict the solution space so as to exclude improper solutions or to simply fix the offending parameter estimate to have a value on border of the permissible solution space (e.g., when variance estimates are negative they can be fixed at zero or at a small positive value). These strategies, however, merely make the problem less obvious and rarely have any substantive effect on the results (see Marsh, 1988). Heywood cases may also be symptomatic of poor models, particularly when parameter estimates are far outside of the range of permissible values or when the standard errors for the offending parameters are very large. The problem, of course, is how to determine whether Heywood cases are due to a poor model or to sampling fluctuations. Dillon, Kumar and Mulani (1987, p.134) offered the following advice: "if the model provides a reasonable fit, the resper .ve confidence interval for the offending estimate covers zero, and the magnitude of the standard error is roughly the same as the other estimated standard errors, the Heywood case is likely to be due to sampling fluctuations." To this good advise might be added the suggestion that the results are substantively reasonable. Even though Heywood cases are common in CFA studies, their occurrence should always be noted and should dictate caution in subsequent interpretations.

Results

MF Factors Inferred Across All Five MF Instruments.

The first-order factor model. MF factors described here are based on all 30 MF parcels representing the five MF instruments (i.e., 3 M and 3 F parcels for each MF instrument). Model 1 (Tables 2 and 3, considering only the total group (TG) analyses for now) is a first-order model. It explains responses to the 30 parcels in terms of 10 first-order factors -- an M and an F factor for each of the five instruments. The factor structure is well defined in that all factor loadings are statistically significant, each of the 10 factors accounts for a significant portion of the variance, and the fit of Model 1 (Table 4) is reasonable. This first-order model is important

because its goodness of fit establishes an upper-limit for the higher-order models based on the same data (i.e., the models are nested) and because higher-order models are based on it. The purpose of higher-order models is to describe correlations among the first-order factors in terms of higher-order factors, and so the correlations among the 10 first-order factors in Model 1 (Table 3) are particularly important.

Insert Tables 2, 3 & 4 About Here

The MTMM perspective. The 10 first-order factors in Model 1 correspond to M and F traits inferred from each of five MF instruments. The correlations among these first-order factors (Table 3) represent a MTMM matrix in which the multiple traits are M and F, and the multiple methods are the five MF instruments. In MTMM studies it is typical to assess convergent validity, discriminant validity, and method/halo effects. Convergent validity is agreement between measures of the same trait assessed by different methods. In MTMM terminology, the 10 correlations among the M factors (.46 to .98; median = .59) and the 10 correlations among the F factors (.23 to .80; median = .64) are convergent validity coefficients. Discriminant validity refers to the distinctiveness of the different traits, the ability to distinguish M from F. Method/halo effects are undesirable biases that are idiosyncratic to a particular method of measurement. Because the five MF instruments were constructed differently, particularly with regard to the social desirability of items, it is likely that method effects do exist and that these method effects are related to social desirability. For example, BSRI and the PAQ instruments contain only socially desirable characteristics, and so it is likely that correlations between M and F will be biased by social desirability when based on these instruments.

MTMM matrices have traditionally been examined according to guidelines such as those developed by Campbell and Fiske (1959; also see Marsh, in press). Whereas the guidelines are useful (Marsh, in press), they have been criticized and many researchers advocate the use of CFA for MTMM data (e.g, Bagozzi, 1980; Joreskog, 1974; Kenny, 1979; Marsh, in press; Marsh & Hocevar, 1983; Schmitt & Stults, 1986; Widaman, 1985). In the CFA approach factors defined by the multiple indicators of the same trait support the construct validity of traits, whereas factors defined by variables representing the same method argue for method/halo effects. These researchers typically recommend that there should be at least three traits and three methods so that each factor is defined by at least three measured variables. In the present investigation there are only two traits, but Kenny

(1979; also see Marsh, 1988; in press) described an alternative parameterization of the MTMM model for this situation that is used here. In this parameterization, method variance is inferred from correlated residuals for variables that share the same method of measurement (see Appendix I). Marsh (1988; in press) examined this alternative parameterization of method effects and recommended it for all MTMM studies even when there are three or more traits and methods.

In most applications of CFA to MTMM data (e.g., Widaman, 1985) trait and method effects are inferred on the basis of correlations among scale scores that represent each trait/method combination. This could be accomplished here by taking an average of the parcels (or, equivalently, the original items) used to define the M and F scores for each instrument, and using these 10 scale scores as the starting point of subsequent analyses. This 10x10 matrix of correlations among scale scores would be similar in many respects to the corresponding matrix of correlations among the 10 first-order factors (i.e., latent constructs) in Table 3. The two correlation matrices would differ in that: (a) the latent constructs are optimally weighted combinations of measured variables whereas corresponding scale scores are not; (b) the latent constructs are corrected for measurement error whereas the corresponding scale scores are not; and (c) the fit of the model used to derive the latent constructs (i.e., Model 1) is explicitly tested as part of the analysis whereas the implicit factor structure used to compute the scale scores is typically untested. Marsh and Hocevar (1988) noted these advantages and argued that it is better to infer trait and method effects on the basis of correlations among latent traits instead of correlations among scale scores. They described how this could be accomplished with the use of HCFA.

The HCFA approach to MTMM data. Conceptually, a second-order factor analyses is like conducting two separate factor analyses. The first factor analysis is performed on relations among measured variables (item or parcel scores) to obtain first-order factors. The second factor analysis is performed on relations among the first-order factors to obtain second-order factors. In the HCFA approach to higher-order factor analysis, both the first- and second-order factors are actually estimated simultaneously. As already noted, however, it is useful to carefully examine the fit and parameter estimates for the first-order model before proceeding to the higher-order models. This approach to HCFA is described in greater detail by Marsh (1985; 1987a; 1987b; Marsh & Hocevar, 1985) and applied to MTMM data

by Marsh and Hocevar (1988).

In the HCFA approach to MTMM data (Marsh & Hocevar, 1988) each trait/method combination is represented by a latent construct, one of the first-order factors in Model 1. Trait and method effects are inferred on the basis of second-order factors. Marsh and Hocevar (1988) described how the models typically used to test for these effects in the CFA of MTMM data (e.g., Marsh, in press; Widaman, 1985) can easily be translated into second-order models so long as there are multiple indicators of each trait/method combination.

The second-order factor models considered here are illustrated in Figure 1 (also see Appendix I). In various models the 5 first-order M factors are used to define a second-order global M factor (GM in Models 3, 4, 5 and 6), the 5 first-order F factors are used to define a second-order global F (GF in Models 3,4,5 and 6) factor, and all 10 first-order factors are used to define a global trait factor (GMF in Models 2, 5 and 6). An essential difference between these models is the number of higher-order factors that are hypothesized.

In Models 4 and 6 method effects are tested by allowing correlations between the residual variance estimates (variance unexplained in terms of higher-order factors) of first-order factors derived from the same MF instrument. That is, a method effect is inferred when the correlation between two different traits (M and F) derived by the same method (instrument) is idiosyncratic to that method. For example, if the BSRI M and the BSRI F scores are more highly correlated than can be explained in terms of the correlation between global M and global F, a method effect is inferred. This representation of method effects is particularly useful when there are only two traits associated with each method of measurement (Kenny, 1979; Marsh, in press). When each method of measurement is represented by at least three traits, method effects can also be represented as method factors (see Marsh, in press for a comparison of the two approaches).

In HCFA, relations among first-order factors are fixed to be zero and these relations are represented in terms of higher-order factors. For example, Model 2 posits that all the relations among the first-order factors (Table 3) can be explained in terms of just one second-order factor (GMF). Because Model 2 has fewer estimated parameters, it is more parsimonious than the corresponding first-order Model 1. It is important to emphasize that Models 2 - 6 positing higher-order factors are all nested under the first-order model (Model 1) so that none can fit the data any better than Model 1. In this respect the fit of the first-order model represents an optimum or

target for the fit of all the higher-order models.[1] The higher-order models are, however, more parsimonious in that they use fewer parameters to fit the data. Thus, to the extent that the fit of a higher-order model approaches that of the corresponding first-order model and the parameter estimates support the posited constructs, then there is support for the higher-order model.

Inferring global M (GM), global F (GF), and global MF (GMF) factors. HCFA models in Figure 1 (also see Appendix I) posit second-order trait factors (GM, GF, GMF) and method effects (correlated residuals) to explain correlations among the first-order factors. For now, only models fitted to the total group covariance are considered. The ability of alternative models to fit the data and their parameter estimates are used to infer the existence of trait and method effects. In Model 2 (Figure 1A) a single higher-order factor is posited to account for all the covariation among the 10 first-order factors. If first-order M and F factors consistently load in the opposite direction on the second-order (GMF) factor and Model 2 is able to fit the data, then the results would support the bipolarity of MF. Inspection of the higher-order factor loadings (not shown) demonstrated that this factor was bipolar, but the model fits (TLI=.693 in Table 4) the data more poorly than models positing two or three higher-order factors. Much of the covariation among first order factors is unexplained by global GMF.

In Model 3 (Figure 1B) two higher-order factors, GM and GF, are posited. This two-factor model provides a better fit (TLI=.755) than the one-factor model. Also, the modest correlation between GF and GM (-.23) indicates that GF and GM are distinguishable (i.e., not bipolar). In Model 4 (Figure 1C), five correlated residuals are added to the Model 3 to test for method effects. The inclusion of the correlated residuals substantially improved the fit (TLI=.805), implying that there are method effects. Furthermore, the correlation between GF and GM in Model 4 (-.36) is more negative than in Model 3 (-.23). This suggests that the method variance may have influenced the earlier estimate of the GM/GF correlation in Model 3.

Model 5 (Figure 1D) combines the GMF factor posited in Model 2 and the GM and GF factors posited in Model 3. In Model 5 correlated residuals are not posited. Model 5 provided three well-defined higher-order factors and produced a substantially improved fit (TLI=.850). In Model 6 (Figure 1E), the five correlated residuals used to infer method effects were added and the fit improved modestly. The TLI (.866) for Model 6 is reasonable and the same as that of Model 1, indicating that most of the covariation among the first-order factors in Model 1 can be explained by Model 6. Because Model 6

requires 19 fewer parameter estimates than Model 1, Model 6 is more parsimonious than Model 1.

A more detailed inspection of parameter estimates in Model 6 (Table 5) facilitates the interpretation of these higher-order factors. Four of five M factors -- all but the CPS -- load positively and significantly on GM; three of five F factors -- all but CPS and CPI -- load positively and significantly on GF; all 5 M factors load significantly and positively on GMF and all 5 F factors load significantly and negatively on GMF. Thus the higher-order factors are well-defined.

The CPS M and F factors that earlier analyses showed to represent a bipolar factor load almost exclusively on GMF. The CPI was also designed to measure a bipolar MF, and the CPI M and F factors tend to have higher loadings on GMF than on GM or GF. The PAQ and BSRI were designed to measure distinguishable M and F factors with socially desirable items, and factors from these two instruments tend to have higher loadings on GM and GF than on GMF. The ANDRO M and F factors were also designed to infer distinguishable M and F factors, but they load more substantially on the GMF factor than on the GM and GF factors. However, the ANDRO M and F scales tend to be negatively correlated with social desirability (Table 1). This suggests that the GM and GF factors in Model 6 may reflect primarily the socially desirable aspects of the masculine and feminine stereotypes. Consistent with this interpretation but in contrast to earlier models, the correlation between GM and GF is positive (.34) instead of negative as in Models 2 - 4.

Insert Table 5 About Here

In Model 6 three higher-order trait factors were posited, and correlated residuals were used to assess method effects. The addition of the correlated residuals in this model had a much smaller effect (Model 6 vs. 5) than when only two higher-order trait factors were estimated (Model 4 vs. Model 3). This suggests that much of what initially appeared to be error due to method effects can be explained in terms of the three higher-order trait factors. The results have important theoretical implications in that they provide support for both the bipolar GMF posited by traditional personality theorists and the separate GM and GF factors posited in androgyny theory.

Despite the intuitive appeal of the interpretation of Model 6, there are also problems. First, the correlation between the residuals for the CPS M and F factors is larger than the residual variance for either factor. This problem was demonstrated in Table 1 when the correlation between M and F was more negative than -1.0 after correction for attenuation, and was also

found in a CFA study based on the normative data base for the CPS instrument (Marsh, 1985). Thus, the problem is not specific to this model or even to this data. Using item-level data, Marsh (1985) demonstrated that this situation was due to the fact that CPS M items were logically opposed to CPS F items. Using opposites forced the correlation between the M and F scores to be more negative than would be expected from the internal consistency of responses within each scale. Second, the residual variance term for the PAQ M factor is slightly negative. Since this offending parameter is not significantly different from zero and its standard error is not excessive, this Heywood case is apparently due to sampling error. [The residual variance is the amount of variance in a first-order factor that is unexplained in terms of second-order factors and small residuals mean that a first-order factor is well-explained by higher-order factors]. These problems, though apparently not serious, dictate caution in interpreting the results.

In summary, three higher-order traits are defined by the set of five MF instruments. One factor is clearly identified as the bipolar GMF posited in traditional personality instruments such as the CPS. However, the reasonably distinguishable facets of GF and GM posited by androgyny theory are also clearly evident. The pattern of loadings and the positive correlation between GM and GF suggest that these higher-order traits are inferred from socially desirable attributes that are relatively unique to masculine and feminine stereotypes, and this interpretation also appears to be consistent with androgyny theory. Further tests of the construct validity of these interpretations will be considered in the next section.

MF Factors: Their Relation to Other Constructs.

The purpose of this section is to examine relations between the higher-order MF factors, social desirability, esteem, and other constructs. This is accomplished by adding measures of these new constructs to models considered in the last section. These relations between the previously identified factors and these new constructs are used to test the construct validity of earlier interpretations of GM, GF and GMF. The nature of these tests is discussed in more detail as part of the presentation of the results. Because the relations between MF responses and these constructs are not the major focus of the present investigation and were examined in detail by Marsh, Antill and Cunningham (1987) using the same data, the results are considered only briefly here.

Insert Table 6 About Here

Social desirability. In the first pair of analyses, a social

desirability factor (inferred from the 6 parcels, 3 from each of the two social desirability measures) was added to the 10 MF factors considered earlier. In one such model, social desirability was related to GM and GF (i.e., a GMF was not posited). As demonstrated in Table 1 for raw scale scores, the social desirability factor was positively correlated with BSRI and PAQ responses, relatively uncorrelated with CPS and ANDRO responses, and negatively correlated with CPI responses. Social desirability was also substantially more positively correlated with GF than with GM (Table 6).

When GM, GF, and GMF were posited, social desirability was more positively correlated with both GM and GF, but relatively uncorrelated with GMF. These observations are consistent with earlier interpretations suggesting that GM and GF represented primarily the socially desirable aspects of M and F when GM, GF and GMF were included in the same model.[2]

Esteem. In the second pair of analyses, an esteem factor (inferred from the 6 parcels, 3 from each of the two esteem measures) was added to the 10 MF factors considered earlier. As demonstrated with the raw scale scores in Table 1, esteem was substantially more positively correlated with M than with F for each of the MF instruments. When just two higher-order factors (GM and GF) were posited (see Table 6), the GM/esteem correlation (.69) was very large and positive whereas the GF/esteem correlation was small and negative (-.14). However, when three higher-order factors were posited, esteem was positively correlated with GM (.55), GF (.29) and GMF (.43). This is consistent with the suggestion that the GM and GF factors reflect socially desirable aspects of M and F.

Insert Table 6 About Here

Biological gender and the adjectives "masculine" and "feminine". In order to further test the construct validity of interpretations of the higher-order MF factors, gender and responses to the adjectives "masculine" and "feminine" (items from BSRI) were added to models with two higher-order (GM, GF) factors and to models with three higher-order (GM, GF, GMF) factors. Biological gender (1=male, 2=female) is a bipolar construct, and other researchers (e.g., Pedhauzur and Tetenbaum, 1979) have reported that the adjectives "masculine" and "feminine" from the BSRI define a two-item bipolar factor. Support for the earlier interpretation requires that each of these new variables should correlate in the appropriate direction with the three higher-order factors, but that each should correlate substantially more with GMF than with either GM or GF.

When just two higher-order (GM, GF) factors are posited, biological

gender, the adjective "masculine", and the adjective "feminine" are each correlated in the expected direction with GM and GF (Table 6). When three higher-order (GM, GF, GMF) factors are considered, Gender (1=male, 2=female) is positively correlated with GF and negatively correlated with GM. Gender, however, is substantially more related to GMF than to either GM or GF. The single-item factor defined by responses to the adjective "masculine" is positively correlated with GM and negatively correlated with GF, but it is much more substantially correlated with GMF. The single-item factor defined by responses to the adjective "feminine" is positively correlated with GF and negatively correlated with GM, but again its largest correlation is with GMF. Because these additional constructs are bipolar constructs and they correlate more substantially with bipolar GMF than with GM or GF, the results support the construct validity of interpretations of the three higher-order factors.

Examination of the MF factor structure within, across, and between gender groups.

Parameter estimates for CFA models can be examined for responses by men and women separately, for the total group covariance matrix, or for the pooled within-group covariance matrix that removes the effect of gender. Because there are substantial gender differences in responses to M and F scales, each approach is likely to result in different parameter estimates. Theoretical, empirical, and pragmatic considerations led to the decision to focus on the total group covariance matrix in the present investigation. The purpose of discussion and results presented in this section is to further examine the basis of this decision and its implications.

Are the factor structures underlying responses to MF responses similar for men and women? A host of theoretical and philosophical issues relate to this question, but the present focus is on methodological issues. Comparing responses by men and women requires at least certain aspects of the factor structures to be equivalent, and pooling responses across groups requires even more stringent assumptions (e.g., Cole & Maxwell, 1985). Whereas exploratory factor analysis is generally inappropriate for examining issues of factorial invariance, CFA is ideally suited to this purpose (see Marsh, 1985). With multigroup CFA, the equivalence of any one or any set of parameter estimates across groups can be tested, and hierarchies of nested models have been proposed for this purpose (e.g., Alwin & Jackson, 1981; Cole & Maxwell, 1985; Joreskog & Sorbom, 1981; Marsh & Hocevar, 1985). The most general test, no matter what the hypothesized model, is a test of the

equality of the entire variance/covariance matrix across groups. This test, Box's M (see Cole & Maxwell, 1985), can be conducted using the MANOVA procedure in SPSSx (1986) which also creates a pooled-within group covariance matrix. The logic of this test is that so long as the covariance matrices are equivalent, then structures based on more restrictive models will also be equivalent.

Two different tests of the equality of covariance matrices based on responses by men and by women were conducted. First, the equivalence of the 30x30 covariance matrices representing the 30 MF parcels derived from all 5 MF instruments was tested. The X^2 of 510 (df=465, N=237, p> .05) was not significant, thus supporting the equivalence of the covariance matrices. Second, the equivalence of the 42x42 covariance matrices representing the 30 MF parcels, the 6 esteem parcels, and the 6 social desirability parcels was tested. Again, the X^2 of 974 (df=903, N=237, p > .05) was not statistically significant, thus supporting the equivalence of these expanded covariance matrices. The omnibus nature of these tests (i.e., the simultaneous test of a large number of parameters), the modest sample sizes, and the use of nonsignificant statistical tests as a basis of support for a null hypothesis all dictate caution in the interpretation of the finding. Nevertheless, the findings provide a reasonable basis for pooling responses by men and women in subsequent analyses.

A second issue is whether analyses should be performed on the pooled within-group covariance matrix that removes the effect of gender, or on the total group covariance matrix that includes the effect of gender. It is well-known that spurious correlations can result when groups differing on some irrelevant variable are combined. However, the effect of gender can hardly be considered an irrelevant variable in the study of MF. To the extent that gender is a valid source of variance to MF responses, then it is theoretically appropriate to conduct analyses on the total group covariance matrix, as in the present investigation. However, because it is also relevant to know how gender affects the MF factor structure, additional analyses were conducted on the pooled within-group covariance matrix for selected models (those results designated by WG in Tables 2-5). The comparison of results of the same model fit to these two matrices -- the total group and the pooled within-group covariance matrices -- indicates the effect of gender on the MF factor structure.

The first-order (model 1) model posited 10 MF factors to fit responses to the 30 MF parcels. This model was fit to both the total-group covariance

matrix (Model 1) and the pooled within-group covariance matrix in which the effect of gender is removed (Model 1a). Whereas the factor structure is well-defined in both analyses (Tables 2-4), there are important differences. The factor loadings are systematically smaller for the pooled within-group analysis (Table 2). Since some of the variance in the M and F factors is related to biological gender, partialling out the effect of gender reduces the variance in M and F factors.

There are also systematic differences in the correlations among the factors for the two analyses (Table 3). Correlations among the different M factors and among the different F factors are smaller for the pooled within-group analyses. Hence, partialling out the effect of gender reduces the apparent agreement among the different MF instruments. Also, correlations between M factors and F factors are less negative in the pooled within-group analyses. Hence, partialling out the effects of gender reduces the apparent "bipolarity" of MF responses.

Selected higher-order models were also fit to the within-group covariance matrix (Table 4). The fit of these models is somewhat better, partly because there is less covariance to be explained when the effect of gender is removed. The comparison of the relative fit of the models again supports the inclusion of GM, GF, GMF to represent trait effects and correlated errors to represent method effects (Model 6a). When just GM and GF are posited, the negative GM/GF correlation observed in Model 4 is close to zero when based on the pooled within-group covariance matrix (Model 4a). When GM, GF, and GMF factors are posited, factor loadings for GM and GF are little affected, but factor loadings for GMF are generally smaller for the analysis of the pooled within-group matrix (Table 5). These results suggest that although factor structures are similar for both analyses, removing the effect of gender reduces the apparent bipolarity of MF responses.

The observed pattern of differences between analyses based on total-group and pooled within-group covariance matrices is not surprising. In fact, the construct validity of the MF responses would be suspect if such a pattern had not occurred. The results do, however, demonstrate the implications of this important methodological consideration. As described earlier, the theoretical position taken here is that the total group analyses are appropriate because gender is a valid source of variance to MF responses. From this perspective, results presented here support the construct validity of the MF responses.

The finding that separate covariance matrices based on responses by men

and by women are not significantly different has important methodological implications that were emphasized here. The substantive implications of these findings may, however, be even more important. First, the finding implies that the factor structure underlying responses to a diverse set of MF instruments do not differ significantly for men and women. Second, the finding implies that the relations of MF responses to esteem and social desirability do not differ significantly for men and women.

Discussion

The Structure of MF Responses

The most salient distinction between androgyny and traditional approaches to the study of MF has been the proposed structure of MF. Previous research has focused on the choice between distinguishable M and F traits posited by androgyny theory, and the single bipolar MF trait posited by traditional personality approaches as if the two models were mutually exclusive. It is clear, however, that MF instruments can be constructed so as to produce either bipolar or relatively independent traits. For example, the M and F traits measured by the BSRI and PAQ may be more accurately designated as measures of assertiveness/dominance and of nurturance respectively (Spence, 1984), and these traits are relatively independent. Furthermore, the use of just socially desirable items on the BSRI and PAQ is likely to produce MF correlations that are more positive than scales that are balanced in relation to social desirability. In contrast, items strongly linked to gender (as on the CPI) or logically opposed items (as on the CPS) will produce a much more negative MF correlation. From this perspective an important substantive contribution of the present investigation is the demonstration that three higher-order MF factors are needed to explain responses to the five MF instruments. In contrast to previous demonstrations that sought to contrast one-factor (GMF) and two-factor (GM and GF) structures, the present results clearly identified all three (GM, GF, and GMF) factors. Thus the results provide support for both the androgyny and the traditional perspectives.

The idea that GM, GF, and GMF all exist simultaneously may be novel, but the empirical support for this contention has been found previously. Three orthogonal factors similar to the ones found here were reported by Pedhauzer and Tetenbaum (1979) in their factor analysis of BSRI responses. Their M and F factors were defined by socially desirable masculine and feminine characteristics whereas their bipolar MF factor was defined by the adjectives "masculine" and "feminine." Since the "masculine" and "feminine" adjectives

had such strong face validity they interpreted the findings to mean that the M and F factors may lack validity. Instead, the present results suggest that all three factors represent distinct components of the MF construct.

The Relation of MF to Gender.

Gender is consistently related to M in one direction and to F in the opposite direction. Thus it is no surprise that removing the effect of gender reduces the variance in both M and F, and also makes the MF correlation less negative or more positive. The position taken here is that this variance attributable to gender is a valid source of variance in MF responses. Perhaps a more neutral position is that it is a source of variance that needs to be considered. This is relevant, because variance attributable to gender is typically eliminated when researchers conduct separate analyses on responses by men and by women. The decision to use separate group covariance matrices is often based on assumed differences in the factor structure for responses by men and women. However, the present investigation provided support for the invariance of the factor structures, and this is a substantively important finding. Nevertheless, it must be emphasized that even when within-group factor structures are equivalent, this within-group factor structure will differ systematically from the total group factor structure.

This methodological issue also has important implications for other personality research that examines factor structures within, across or between subgroups that are not amenable to random assignment (e.g., sex, race, SES, age). Typically there is no a priori basis for concluding that any one approach is necessarily superior. As demonstrated here, the best approach is to compare the empirical and theoretical implications of the different approaches. In pursuing this comparison, an omnibus test of the equality of subgroup covariance matrices such as Box's M is a useful starting point. When there is support for this equality, subsequent analyses of either the total-group or pooled within-group covariance matrix is justified and the comparison of both approaches is recommended. When the omnibus test indicates that the subgroup covariance matrices are not equivalent, further analyses can be conducted to determine what aspects (e.g., factor loadings, factor correlations, uniquenesses) of the first-order or second-order factor structures differ (see Alwin & Jackson, 1981; Marsh & Hocevar, 1985).

Footnotes

[1] Marsh (1987a) and Marsh and Hocevar (1985) use this relation between a higher-order factor model and its corresponding first-order factor model to define the target coefficient that is used as a goodness-of-fit indicator in Table 4.

[2] In supplemental analyses, the effects of social desirability were partialled out of MF responses (see Appendix I). This made MF correlations more negative for PAQ, BSRI and CPI, but had almost no effect for ANDRO and CPS. Correlations among M factors, and correlations among F factors, were somewhat higher when the effect of social desirability was removed. Because these correlations are the convergent validities in MTMM analyses, these results are consistent with earlier suggestions that social desirability acts like a method effect. Partialling out social desirability also substantially reduced the effect of introducing correlated uniquenesses. Thus, much, but apparently not all, of the method effects were associated with the social desirability factor.

References

Alwin, D. F., & Jackson, D. J. (1981). Applications of simultaneous factor analysis to issues of factorial invariance. In D. J. Jackson & E. F. Borgotta (Eds.), Factor analysis and measurement in sociological research. Beverly Hills, CA: Sage.

Anderson, J. C., & Gerbing, D. W. (1984). The effect of sampling error on convergence, improper solutions, and goodness-of-fit indices for maximum likelihood confirmatory factor analysis. Psychometrika, 49, 155-173.

Antill, J. K. & Cunningham, J. D. (1979). Self-esteem as a function of masculinity in both sexes. Journal of Consulting and Clinical Psychology, 47, 783-785.

Antill, J. K. & Cunningham, J. D. (1980). The relationship of masculinity, femininity, and androgyny to self-esteem. Australian Journal of Psychology, 32, 195-207.

Bagozzi, R. P. (1980). Causal models in marketing. New York: Wiley.

Bem, S. L. (1974). The measurement of psychological androgyny. Journal of Consulting and Clinical Psychology, 42, 155-162.

Bentler, P. M. & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. Psychological Bulletin, 88, 588-606.

Berger, E. M. (1952). The relation between expressed acceptance of self and expressed acceptance of others. Journal of Abnormal and Social Psychology, 47, 778-782.

Berzins, J. I., Welling, M.A., & Wetter, R.E. (1978). A new measure of psychological androgyny based on the Personality Research Form. Journal of Consulting and Clinical Psychology, 46, 126-138.

Campbell, D. T. and Fiske, D. W. (1959). Convergent and discriminant validation by multitrait-multimethod matrix. Psychological Bulletin, 56, 81-105.

Cole, D. A., & Maxwell, S. E. (1985). Multitrait-multimethod comparisons across populations: A confirmatory factor analytic approach. Multivariate Behavioral Research, 20, 389-417.

Comrey, A. L. (1970). Comrey Personality Scales manual. San Diego: CA: Educational and Industrial Testing Service.

Cook, E. P. (1985). Psychological androgyny. New York: Pergamon.

Constantinople, A. (1973). Masculinity-femininity: An exception to a famous dictum? Psychological Bulletin, 80, 389-407.

Dillon, W. R., Kumar, A., & Mulani, N. (1987). Offending estimates in

covariance stgructure analysis: Comments on the causes of and the solutions to Heywood cases. Psychological Bulletin, 101, 126-135.

Eagly, A.H. (1967). Involvement as a determinant of response to favorable and unfavorable information. Journal of Personality and Social Psychology Monograph (3, Whole No. 643).

Gerbing, D.W., & Anderson, J.C. (1987). Improper solutions in the analysis of covariance structures: Their interpretability and a comparison of alternative respecifications. Psychometrika, 52,99-111.

Gough, H. G. (1957). California Psychological Inventory manual. Palo Alto: Consulting Psychologists Press.

Hall, J. A., & Taylor, M. C. (1985). Psychological androgyny and the masculinity x femininity interaction. Journal of Personality and Social Psychology, 49, 429-435.

Jackson, D. (1967). Personality Research Form Manual. Goshen, NY: Research Psychologist Press.

Janis, I.L., & Fields, P.B. (1959). Sex differences and personality factors related to persuasibility. In C.I. Hovland & I.L. Janis (Eds.) Personality and persuasibility (p. 55-68). New Haven, CN: Yale University Press.

Joreskog, K. G. (1981). Analysis of covariance structures. Scandinavian Journal of Statistics, 8, 65-92.

Joreskog, K. G. (1974). Analyzing psychological data by structural analysis of covariance matrices. In R. C. Atkinson, D. H. Krantz, R.D. Luce, & Suppes (Eds.), Contemporary developments in mathematical psychology (Vol. 2, pp. 1-56). San Francisco: W. H. Freeman.

Joreskog, K. G. & Sorbom, D. (1981). LISREL V: Analysis of Linear Structural Relations By the Method of Maximum Likelihood. Chicago: International Educational Services.

Kelly, J., & Worrell, J. L. (1977). New formulations of sex roles and androgyny: A critical review. Journal of Consulting and Clinical Psychology, 45, 1101-1115.

Kenny, D. A. (1979). Correlation and causality. New York: Wiley.

Locksley, A., & Colten, M. E. (1979). Psychological androgyny: A case of mistaken identity? Journal of Personality and Social Psychology, 37 1017-1031.

Long, K. S. (1983) Confirmatory factor analysis: A preface to LISREL. Beverly Hills, CA: Sage.

Lubinski, D., Tellegen, A., & Butcher, J. N. (1983). Masculinity, femininity, and androgyny viewed and assessed as distinct concepts.

Journal of Personality and Social Psychology, 44, 428-439.

Marsh, H. W. (1985). The structure of masculinity/femininity: An application of confirmatory factor analysis to higher-order factor structures and factorial invariance. Multivariate Behavioral Research, 20, 427-449.

Marsh, H. W. (1987a). The hierarchical structure of self-concept and the application of hierarchical confirmatory factor analysis. Journal of Educational Measurement, 24 17-39.

Marsh, H. W. (1987b). Masculinity, femininity and androgyny: Their relations with multiple dimensions of self-concept. Multivariate Behavioral Research, 22 91-118.

Marsh, H. W. (1988). Confirmatory Factor Analyses of Multitrait-multimethod data: Many problems and a few solutions. Manuscript submitted for publication.

Marsh, H. W. (in press). Multitrait-multimethod analyses. In J. P. Keeves (Ed.), The international handbook of educational research methodology, measurement and evaluation. New York: Pergamon Press.

Marsh, H. W., Antill, J. K. & Cunningham, J. D. (1987). Masculinity, femininity, and androgyny: Relations to self esteem and social desirability. Journal of Personality, 55, 661-685.

Marsh, H. W., Balla, J. R. & McDonald, R. P. (in press). Goodness-of-fit indices in confirmatory factor analysis: The effect of sample size. Psychological Bulletin, .

Marsh, H. W., & Hocevar, D. (1983). Confirmatory factor analysis of multitrait-multimethod matrices. Journal of Educational Measurement, 20, 231-248.

Marsh, H. W. & Hocevar, D. (1984). The factorial invariance of students' evaluations of college teaching. American Educational Research Journal, 21, 341-366.

Marsh, H. W. & Hocevar, D. (1985). The application of confirmatory factor analysis to the study of self-concept: First and higher order factor structures and their invariance across age groups. Psychological Bulletin, 97, 562-582.

Marsh, H. W. & Hocevar, D. (1988). A new, more powerful approach to multitrait-multimethod analyses: An application of second-order confirmatory factor analysis. Journal of Applied Psychology, xxx-xxx.

Marsh, H. W., & Myers, M. R. (1986). Masculinity, femininity, and androgyny: A methodological and theoretical critique. Sex Roles, 14, 397-430.

Megargee, E. I. (1972). The California Psychological Inventory Handbook. San Francisco: Jossey-Bass.

Pedhauzer, E. J. (1982). Multiple regression in behavioral research (2nd ed.) New York:Holt, Rinehart & Winston.

Pedhauzer, E. J. & Tetenbaum, T. J. (1979). The Bem Sex-Role Inventory: A theoretical and methodological critique. Journal of Personality and Social Psychology, 37, 996-1016.

Schmitt, N., & Stults, D. M. (1986). Methodological review: Analysis of multitrait-multimethod matrices. Applied Psychological Measurement, 10, 1-22.

Spence, J. T. (1984). Masculinity, femininity, and gender-related traits: A conceptual analysis and critique of current research (p. 1-97). In B.A. Maher & W. B. Maher (Eds.), Progress in Experimental Personality Research (Vol. 13). New York:NY, Academic Press.

SPSS. (1986). SPSSx User's Guide. New York, NY: McGraw-Hill.

Van Driel, O. P. (1978). On various improper solutions in maximum likelihood factor analysis. Psychometrika, 43, 225-243.

Whitley, B. E. (1983). Sex role orientation and self-esteem: A critical meta-analytic review. Journal of Personality and Social Psychology, 44, 765-778.

Widaman, K. F. (1985). Hierarchically nested covariance structure models for multitrait-multimethod data. Applied Psychological Measurement, 9, 1-26.

Table 1

Psychometric Properties of the Masculinity (M), Femininity (F), Self-Esteem (SE) and Social Desirability (Soc Des) Scales

Scale[a]		Alpha	Correlations With: Sex (1=male 2=female)	Soc Des	Esteem	Correlation[b] Between M & F	
MF Scales							
BSRI	M	.88	-.39**	.21*	.50**		
	F	.80	.45**	.27**	-.14*		
	MF	.86	-.54**	.00	.45**	-.16*	(-.19)
CPI	M	.69	-.50**	-.15*	.01		
	F	.63	.41**	-.08	-.40**		
	MF	.66	-.63**	-.04	.29**	-.05	(-.12)
PAQ	M	.87	-.25**	.34**	.64**		
	F	.86	.27**	.45*	.14*		
	MF	.81	-.41*	-.05	.42**	.17*	(+.23)
ANDRO	M	.77	-.38**	.06	.50**		
	F	.65	.32**	.22*	-.23**		
	MF	.78	-.44**	-.08	.47**	-.31**	(-.44)
CPS	M	.71	-.51**	-.05	.29**		
	F	.77	.50**	.05	-.35**		
	MF	.86	-.53**	-.05	.34**	-.77**	(-1.04)
Esteem Scales							
Berger		.90	.00	.46**	.93**		
Janis-Field		.90	-.20**	.37**	.93**		
Esteem Total		.94	-.11*	.43*	1.00		
Social Desirability Scales							
BSRI		.71	.27**	.81**	.40**		
PRF		.51	-.01	.81**	.30**		
Soc Des Total		.75	.16**	1.00	.43**		

Note. Each scale score is the unweighted mean response to items designed to infer that scale. All MF scales were defined as M - F. The esteem and social desirability totals were each the unweighted sum of the two scales (after each was standardized, mean = 0, SD =1.0) used to measure each construct.

* $p < .05$; ** $p < .01$

[a] BSRI = Bem Sex Role Inventory; CPI = California Psychological Inventory; PAQ= Personal Attributes Questionnaire; CPS = Comrey Personality Scales; PRF = Personality Research Form. [b] Correlations in parentheses are corrected for unreliability using the coefficient alpha estimates. Note that after correction for unreliability, the correlation between the CPS M and F scales is slightly more negative than -1.0 which is not possible.

Table 2

The First-order Model For Total Group (TG) and Pooled Within-group (WG) Covariance Matrices (Models 1 and 1a in Table 4): Factor Loadings and Error/Uniquenesses

Factor Variable[a]		First-order Factor Loadings		Error/ Uniquenesses[b]	
		TG	WG	TG	WG
M1	BSRIM1	.74[c]	.63[c]	.21*	.18*
	BSRIM2	.66*	.63*	.21*	.20*
	BSRIM3	.67*	.64*	.14*	.14*
F1	BSRIF1	.42[c]	.36[c]	.23*	.23*
	BSRIF2	.73*	.59*	.07*	.07*
	BSRIF3	.37*	.38*	.19*	.17*
M2	CPIM1	.29c	.26c	.17*	.16*
	CPIM2	.39*	.29*	.16*	.17*
	CPIM3	.34*	.26*	.20*	.21*
F2	CPIF1	.31c	.23c	.18*	.19*
	CPIF2	.31*	.27*	.19*	.19*
	CPIF3	.48*	.47*	.21*	.19*
M3	PAQM1	.64[c]	.62[c]	.17*	.17*
	PAQM2	.70*	.68*	.18*	.18*
	PAQM3	.59*	.56*	.20*	.20*
F3	PAQF1	.43[c]	.39[c]	.31*	.29*
	PAQF2	.62*	.61*	.12*	.12*
	PAQF3	.66*	.64*	.10*	.11*
M4	ANDROM1	.14[c]	.13[c]	.02*	.02*
	ANDROM2	.14*	.12*	.02*	.02*
	ANDROM3	.18*	.17*	.02*	.02*
F4	ANDROF1	.12[c]	.12[c]	.02*	.02*
	ANDROF2	.11*	.11*	.02*	.01*
	ANDROF3	.14*	.11*	.02*	.02*
M5	CPSM1	.67[c]	.55[c]	.69*	.69*
	CPSM2	.80*	.66*	.64*	.64*
	CPSM3	.69*	.48*	.82*	.81*
F5	CPSF1	.85[c]	.77[c]	.77*	.75*
	CPSF2	.76*	.55*	.66*	.65*
	CPSF3	.80*	.62*	.84*	.82*

Note. Parameter estimates are in standardized form to facilitate interpretation. Factor correlations are presented in Table 3.

* $p < .05$.

[a] The measured variables were three randomly formed subscales from each M and F scale (e.g., BSRIM1, BSRIM2, BSRMF3 are the three M subscales from the BSRI that define M1). Because each subscale was allowed to define only one factor, factor loadings are presented as a single column instead of as a 30 (measured variables) by 10 (factors) matrix.

[b] Error/uniquenesses were estimated in a diagonal 30 (variables) x 30 matrix that assumed uncorrelated errors among the variables, and so are presented as a column.

[c] The first factor loading for each factor was fixed at 1.0 to serve as a reference indicator and so no test of statistical significance was performed.

Table 3

The First-order Model For Total Group (TG) and Pooled Within-group (WG) Covariance Matrices (Models 1 and 1a in Table 5): Factor Correlations

		BSRI		CPI		PAQ		ANDRO		CPS	
		M1	F1	M2	F2	M3	F3	M4	F4	M5	F5
M1	TG	1a									
	WG	1									
F1	TG	-.18*	1								
	WG	.05	1								
M2	TG	.52*	-.27*	1							
	WG	.37*	.12	1							
F2	TG	-.42*	.45*	-.09	1						
	WG	-.29*	.22*	.30*	1						
M3	TG	.97*	-.12	.49*	-.42*	1					
	WG	-.97*	.03	.41*	-.36*	1					
F3	TG	.07	.80*	-.10	.23*	.24*	1				
	WG	.19*	.81*	.07	.12	.33*	1				
M4	TG	.81*	-.44*	.59*	-.58*	.87*	-.12	1			
	WG	.76*	-.26*	.43*	-.47*	.86*	-.02	1			
F4	TG	-.26*	.71*	-.21*	.65*	-.24*	.60*	-.43*	1		
	WG	-.11	.62*	.05	.54*	-.14	.58*	-.30*	1		
M5	TG	.46*	-.55*	.60*	-.69*	.52*	-.22*	.72*	-.54*	1	
	WG	.30*	-.33*	.32*	-.56*	.47*	-.09	.63*	-.41*	1	
F5	TG	-.42*	.61*	-.35*	.74*	-.47*	.24*	-.63*	.63*	-1.11*	1
	WG	.24*	.40*	.07	.60*	-.39*	.12	-.51*	.50*	-1.16*	1

Note. See footnotes in Table 2.

* $p < .05$.

[a] In unstandardized form the factor variances were: (.62, .22, .14, .17, .45, .35, .03, .02, .68, .85) for TG and (.41, .13, .07, .06, .38, .15, .02, .01, .31, .59) for WG. All factor variances were statistically significant.

Table 4

MTMM Models Positing Global Trait Factors or Method Effects To Explain Responses To Five MF Instruments

Model	χ^2	df	χ^2/df	TLI[a]	TC[b]	GM/GF Correlation
Total Group Analysis						
0	4249.3	435	9.77	---	---	---
1	783.2	360	2.18	.866	1.000	---
2	1459.6	395	3.70	.693	.537	---
3	1241.2	394	3.15	.755	.631	-.23*
4	1054.8	389	2.71	.805	.743	-.36*
5	890.6	384	2.32	.850	.879	.35*
6	825.2	379	2.18	.866	.949	.34*
Pooled Within-Group Analysis						
0a	3620.0	435	8.32	---	---	---
1a	716.8	360	1.99	.865	1.000	---
4a	907.5	389	2.33	.818	.790	.02
6a	822.3	379	2.17	.840	.872	.51*

Note. Six substantive models (2.1 - 2.6) were fit to the Total Group Covariance Matrix and three of these models (2.1a, 2.4a, 2.6a) were also fit to the pooled within-group covariance matrix. Parameter estimates for Models 2.6 and 2.6a are presented in Table 7. See Appendix II for a description of the models.

* $p < .05$.

[a] TLI = Tucker Lewis index (Bentler & Bonett, 1980). [b] The Target coefficient (TC), designed specifically for HCFA (see Marsh & Hocevar, 1985), is defined as the ratio of the χ^2 for the first-order model (Model 2.1) and any higher-order model. It provides an estimate of the variance in the first-order model that can be explained by the higher-order model. It has a maximum of 1.0 when all covariation among the first-order models can be explained by higher-order factors.

Table 5

HCFA Models 6 (TG) and 6a (WG): Second-order Factor Loadings, First-order Factor Residuals, and Correlations Between First-Order Factor Residuals

		Second-Order Factor Loadings for:						Second-Order Residuals		Correlated Residuals[a]	
		Global M		Global F		Global MF					
Factor		TG	WG	TG	WG	TG	WG	TG	WG	TG	WG
BSRI	M1	.81*	.84*	0	0	.51*	.44*	.10*	.09		
	F1	0	0	.66*	.77*	-.66*	-.49*	.14*	.18*	-.03	-.01
CPI	M2	.29*	.43*	0	0	.51*	.32	.65*	.71		
	F2	0	0	-.03	-.11	-.79*	-.62*	.37*	.61	.32*	.35
PAQ	M3	.89*	.87*	0	0	.52*	.56*	-.05	-.06		
	F3	0	0	.90*	.90*	-.31*	-.24*	.10	.14	.12*	.10*
ANDRO	M4	.52*	.48*	0	0	.78*	.78*	.12*	.15*		
	F4	0	0	.41*	.45*	-.71*	-.64*	.33*	.38*	.07*	.10
CPS	M5	.05	.04	0	0	.89*	.78*	.21*	.39*		
	F5	0	0	.03	.02	-.86*	-.69*	.26*	.52*	-.34*	-.58*

Note. Factor loadings for the 30 measured variables and their error/uniquenesses are not shown because they are so similar to those for the corresponding first-order models (Table 2). Parameter estimates are in standardized form to facilitate interpretation.

* $p < .05$.

[a] Method effects in the MTMM models were represented as correlated residuals between pairs of M and F factors from the same instrument.

Table 6

Relations of Higher-order MF Factors (GM, GF, and GMF) to Social Desirability, Esteem, Gender, and Responses to the Adjectives "Masculine" and "Feminine"

	Models Containing:				
	2 Higher-order Factors		3 Higher-order Factors		
	GM	GF	GM	GF	GMF
Social Desirability	.16*	.49**	.42**	.58**	.16
Esteem	.69**	-.14*	.52**	.34**	.48**
Gender	-.41**	.56**	-.07	.29**	-.58**
Masculine	.53**	-.56**	.19**	-.27**	.63**
Feminine	-.47**	.64**	-.10	.34**	-.67**

Note. Factor correlations are based on two models like those described earlier except that indicators of additional constructs were added that were correlated with the higher-order MF factors. One model included only two higher-order factors (GM and GF) whereas the second contained three higher-order MF factors (GM, GF, and GMF). In three sets of analyses, the higher-order factors were related to: (a) social desirability, (b) esteem, and (c) gender and responses to the adjectives "masculine" and "feminine."

* $p < .05$.

Appendix I -- HCFA Model Specifications in Terms of LISREL Design Matrices

Below are the LISREL design matrices for Model 6 (Figure 1). In this problem there are 30 measured variables (30 MF subscales called mf1 - mf30), 10 first-order factors (M1 - M5; F1 - F5), and 3 second-order factors (GM, GF, GMF) used to explain relations among the 10 first-order MF factors. The four design matrices contain parameters to be estimated (represented as letters a to g), and parameters with fixed values of either 0 or 1. LAMBDA Y is a 30 (measured variables) x 13 (factors) matrix that contains estimated factor loadings (the "a"s) and factor loadings with fixed values (the 1s) that serve as reference indicators. THETA is a 30x30 matrix of uniquenesses (the "b"s) of the measured variables. THETA is specified as a diagonal matrix indicating that uniqueness are uncorrelated, and thus is presented as a single column of values. BETA is a 134x13 matrix that contains second-order factor loadings (the "c"s). PSI is a 13x13 matrix that contains the residual variances for first-order factors (the "d"s), correlations among residual variances that are used to reflect method effects (the "e"s), second-order factor variances fixed to unity (the 1s), and correlations among second-order factors (the "f"s).

Other HCFA models can be easily represented in terms of the four design matrices. For example, Model 4 (Figure 1) differs from the one presented here only in that no only two higher-order factors (GM and GF) were posited. For this model, LAMBDA Y is a 30x12 matrix (the last column is eliminated), BETA and PSI are 12x12 matrices (the last row and column are eliminated) and THETA remains the same. When the six indicators of esteem were added to Model 6 (see Table 6) LAMBDA Y became a 36x14 (reflecting the 6 additional measured variables and 1 additional factor), BETA and PSI became 14x14 matrices (reflecting the one additional factor), and THETA became a 36x36 matrix (reflecting the one additional factor).

Appendix I continued on next page.

Appendix I continued

	LAMBDA Y													THETA
	M1	F1	M2	F2	M3	F3	M4	F4	M5	F5	GM	GF	GMF	
mf1	1	0	0	0	0	0	0	0	0	0	0	0	0	b
mf2	a	0	0	0	0	0	0	0	0	0	0	0	0	b
mf3	a	0	0	0	0	0	0	0	0	0	0	0	0	b
mf4	0	1	0	0	0	0	0	0	0	0	0	0	0	b
mf5	0	a	0	0	0	0	0	0	0	0	0	0	0	b
mf6	0	a	0	0	0	0	0	0	0	0	0	0	0	b
mf7	0	0	1	0	0	0	0	0	0	0	0	0	0	b
mf8	0	0	a	0	0	0	0	0	0	0	0	0	0	b
mf9	0	0	a	0	0	0	0	0	0	0	0	0	0	b
mf10	0	0	0	1	0	0	0	0	0	0	0	0	0	b
mf11	0	0	0	a	0	0	0	0	0	0	0	0	0	b
mf12	0	0	0	a	0	0	0	0	0	0	0	0	0	b
mf13	0	0	0	0	1	0	0	0	0	0	0	0	0	b
mf14	0	0	0	0	a	0	0	0	0	0	0	0	0	b
mf15	0	0	0	0	a	0	0	0	0	0	0	0	0	b
mf16	0	0	0	0	0	1	0	0	0	0	0	0	0	b
mf17	0	0	0	0	0	a	0	0	0	0	0	0	0	b
mf18	0	0	0	0	0	a	0	0	0	0	0	0	0	b
mf19	0	0	0	0	0	0	1	0	0	0	0	0	0	b
mf20	0	0	0	0	0	0	a	0	0	0	0	0	0	b
mf21	0	0	0	0	0	0	a	0	0	0	0	0	0	b
mf22	0	0	0	0	0	0	0	1	0	0	0	0	0	b
mf23	0	0	0	0	0	0	0	a	0	0	0	0	0	b
mf24	0	0	0	0	0	0	0	a	0	0	0	0	0	b
mf25	0	0	0	0	0	0	0	0	1	0	0	0	0	b
mf26	0	0	0	0	0	0	0	0	a	0	0	0	0	b
mf27	0	0	0	0	0	0	0	0	a	0	0	0	0	b
mf28	0	0	0	0	0	0	0	0	0	1	0	0	0	b
mf29	0	0	0	0	0	0	0	0	0	a	0	0	0	b
mf30	0	0	0	0	0	0	0	0	0	a	0	0	0	b

Appendix I continued on next page

Appendix I continued

BETA

	M1	F1	M2	F2	M3	F3	M4	F4	M5	F5	GM	GF	GMF
M1	0	0	0	0	0	0	0	0	0	0	c	0	c
F1	0	0	0	0	0	0	0	0	0	0	0	c	c
M2	0	0	0	0	0	0	0	0	0	0	c	0	c
F2	0	0	0	0	0	0	0	0	0	0	0	c	c
M3	0	0	0	0	0	0	0	0	0	0	c	0	c
F3	0	0	0	0	0	0	0	0	0	0	0	c	c
M4	0	0	0	0	0	0	0	0	0	0	c	0	c
F4	0	0	0	0	0	0	0	0	0	0	0	c	c
M5	0	0	0	0	0	0	0	0	0	0	c	0	c
F5	0	0	0	0	0	0	0	0	0	0	0	c	c
GM	0	0	0	0	0	0	0	0	0	0	0	0	0
GF	0	0	0	0	0	0	0	0	0	0	0	0	0
GMF	0	0	0	0	0	0	0	0	0	0	0	0	0

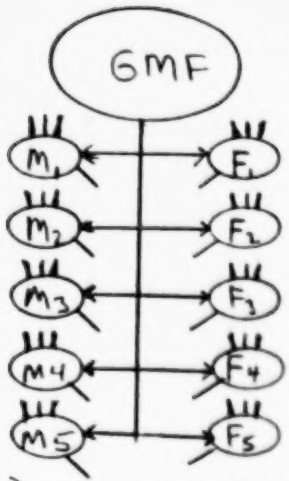
PSI

	M1	F1	M2	F2	M3	F3	M4	F4	M5	F5	GM	GF	GMF
M1	d												
F1	e	d											
M2	0	0	d										
F2	0	0	e	d									
M3	0	0	0	0	d								
F3	0	0	0	0	e	d							
M4	0	0	0	0	0	0	d						
F4	0	0	0	0	0	0	e	d					
M5	0	0	0	0	0	0	0	0	d				
F5	0	0	0	0	0	0	0	0	e	d			
GM	0	0	0	0	0	0	0	0	0	0	1		
GF	0	0	0	0	0	0	0	0	0	0	f	1	
GMF	0	0	0	0	0	0	0	0	0	0	0	0	1

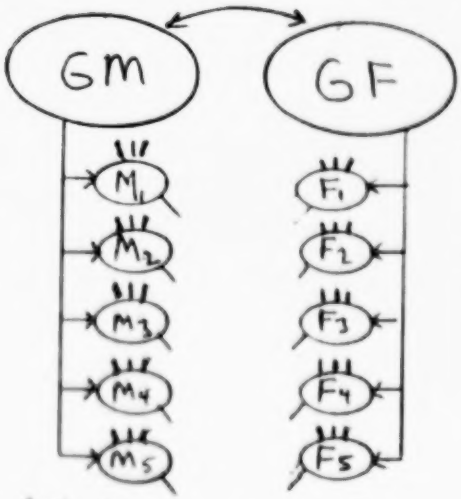
Note. a=first-order factor loadings; b=error/uniquenesses for each measured variable; c=second-order factor loadings; d=first-order factor residuals; e=correlated residuals among first-order factors used to reflect method effects; f=correlations among second-order factors.

Figure Captions

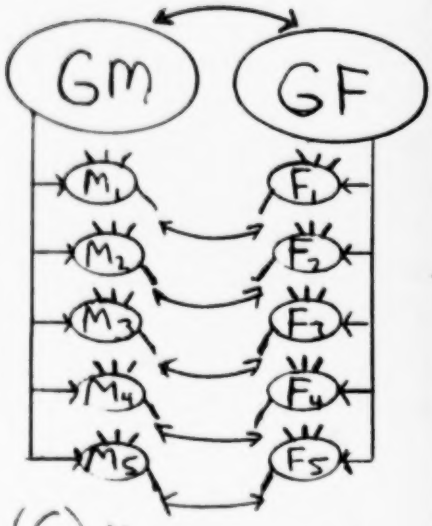
Figure 1. Hierarchical models of the structure of responses to all five masculinity-femininity (MF) instruments. Each of the five models posits one (GMF=global bipolar MF), two (GM=global masculinity and GF=global femininity) or three (GM, GF, and GMF) second-order factors. The second-order factors reflect relations among the first-order M (M1-M5) and F (F1-F5) factors. Each pair of first-order factors (e.g., M1 and F1) represents responses to one of the five MF instruments. (The relations between each first-order factor and its three measured variables are not shown in detail so as to simplify the diagrams.) Two of the models (4 and 6) also contain correlated residuals; these reflect method/halo effects that are idiosyncratic to the pair of first-order factors representing the same MF instrument. The hierarchical structures are presented in terms of LISREL design matrices in Appendix I.



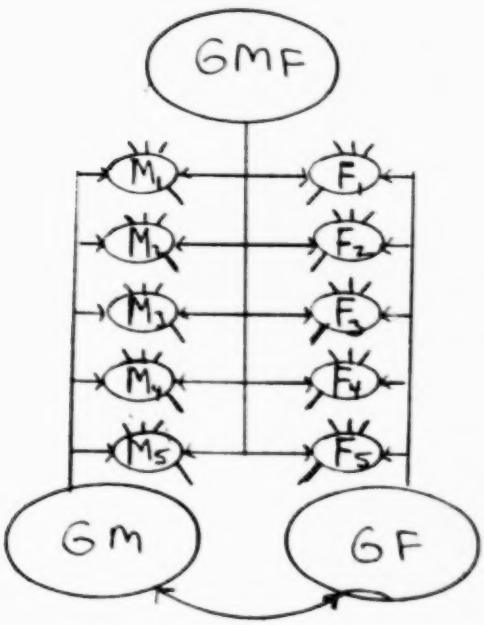
(A) MODEL 2



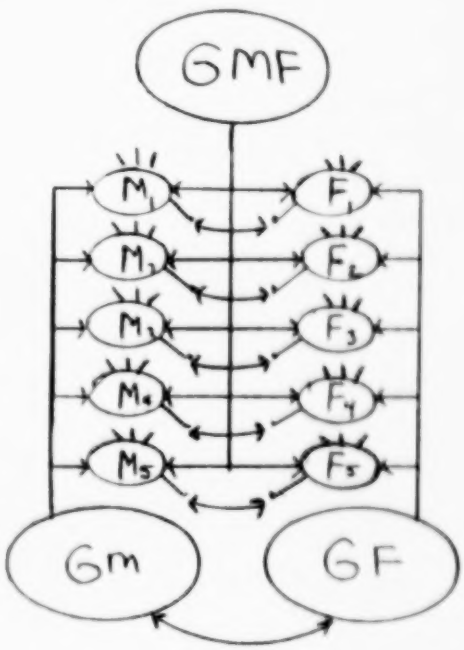
(B) MODEL 3



(C) MODEL 4



(D) MODEL 5



(E) MODEL 6